03.128
YOR

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　５月２１日

出願番号
Application Number: 特願２００３－１４３２２４

[ST. 10/C]: [ＪＰ２００３－１４３２２４]

出願人
Applicant(s): インターナショナル・ビジネス・マシーンズ・コーポレーション

２００３年　９月１９日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０７７５６８

【書類名】　特許願
【整理番号】　JP9030128
【提出日】　平成15年 5月21日
【あて先】　特許庁長官　殿
【国際特許分類】　G10L　3/00
【発明者】
　【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
　【氏名】　滝口　哲也
【発明者】
　【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内
　【氏名】　西村　雅史
【特許出願人】
　【識別番号】　390009531
　【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション
【代理人】
　【識別番号】　100086243
　【弁理士】
　【氏名又は名称】　坂口　博
【代理人】
　【識別番号】　100091568
　【弁理士】
　【氏名又は名称】　市位　嘉宏
【代理人】
　【識別番号】　100108501
　【弁理士】
　【氏名又は名称】　上野　剛史

【復代理人】

【識別番号】　100110607

【弁理士】

【氏名又は名称】　間山　進也

【手数料の表示】

【予納台帳番号】　062651

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0207860

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　音声認識装置、音声認識方法、該音声認識方法をコンピュータに対して実行させるためのコンピュータ実行可能なプログラムおよび記憶媒体

【特許請求の範囲】

【請求項１】　コンピュータを含んで構成され音声を認識するための音声認識装置であって、該音声認識装置は、

音声信号から得られる特徴量をフレームごとに格納する記憶領域と、

音響モデル・データおよび言語モデル・データをそれぞれ格納する格納部と、

その時点で処理するべき音声信号よりも前に取得された音声信号から残響音声モデル・データを生成し、残響音声モデル・データを使用して適合音響モデル・データを生成する残響適合モデル生成部と、

前記特徴量と前記適合音響モデル・データと前記言語モデル・データとを参照して音声信号の音声認識結果を与える認識処理手段と

を含む、音声認識装置。

【請求項２】　前記適合音響モデル生成手段は、ケプストラム音響モデル・データから線形スペクトル音響モデル・データへのモデル・データ領域変換部と、

前記線形スペクトル音響モデル・データに前記残響音声モデル・データを加算して尤度最大を与える残響予測係数を生成する残響予測係数算出部と

を含む、請求項１に記載の音声認識装置。

【請求項３】　さらに残響音声モデル・データを生成する加算部を含み、前記加算部は、前記音響モデルのケプストラム音響モデル・データおよびフレーム内伝達特性のケプストラム音響モデル・データを加算してフレーム内残響影響を受けた音声モデルを生成する、請求項２に記載の音声認識装置。

【請求項４】　前記加算部は、生成された前記フレーム内残響影響を受けた音声モデルを前記モデル・データ領域変換部へと入力し、前記モデル・データ領域変換部に対して前記フレーム内残響影響を受けた音声モデルの線形スペクトル音響モデル・データを生成させる、請求項３に記載の音声認識装置。

【請求項５】　前記残響予測係数算出部は、入力された音声信号から得られた少なくとも１つの音韻と、前記残響音声モデル・データとを使用して線形スペク

トル音響モデル・データに基づいて残響予測係数の尤度を最大化させる、請求項４に記載の音声認識装置。

【請求項６】　前記音声認識装置は、隠れマルコフ・モデルを使用して音声認識を実行する、請求項５に記載の音声認識装置。

【請求項７】　コンピュータを含んで構成され音声を認識するための音声認識装置に対して音声認識を実行させるための方法であって、前記方法は、前記音声認識装置に対して、

音声信号から得られる特徴量をフレームごとに記憶領域に格納させるステップと、

その時点で処理するべき音声信号よりも前に取得された音声信号を前記格納部から読み出して残響音声モデル・データを生成し、格納部に格納された音響モデルを処理して適合音響モデル・データを生成して記憶領域に格納させるステップと、

前記特徴量と前記適合音響モデル・データと格納部に格納された言語モデル・データとを読み込んで音声信号の音声認識結果を生成させるステップと

を含む、音声認識方法。

【請求項８】　前記適合音響モデル・データを生成するステップは、加算部により前記読み出された音声信号とフレーム内伝達特性値との合計値を算出するステップと、

前記加算部により算出された前記合計値をモデル・データ領域変換部に読み込ませ、ケプストラム音響モデル・データから線形スペクトル音響モデル・データへと変換させるステップと、を含む、請求項７に記載の音声認識方法。

【請求項９】　加算部に対して前記線形スペクトル音響モデル・データと前記残響音声モデル・データとを読み込ませ加算して、尤度最大を与える残響予測係数を生成させるステップと

を含む、請求項８に記載の音声認識方法。

【請求項１０】　前記線形スペクトル音響モデル・データへと変換させるステップは、前記加算部に対して、前記音響モデルのケプストラム音響モデル・データおよびフレーム内伝達特性のケプストラム音響モデル・データを加算してフレ

ーム内残響影響を受けた音声モデルを生成するステップを含む、請求項９に記載の音声認識方法。

【請求項１１】　前記残響予測係数を生成させるステップは、前記加算部により生成された前記フレーム内残響影響を受けた音声モデルの線形スペクトル音響モデル・データと前記残響音声モデル・データとの合計値が音声信号から生成され格納された少なくとも１つの音韻に対して最大の尤度を与えるように残響予測係数を決定するステップを含む、請求項１０に記載の音声認識装置。

【請求項１２】　請求項７から請求項１１のいずれか１項に記載された音声認識方法をコンピュータに対して実行させるためのコンピュータ可読なプログラム。

【請求項１３】　請求項７から請求項１１のいずれか１項に記載された音声認識方法をコンピュータに対して実行させるためのコンピュータ可読なプログラムを記憶した、コンピュータ可読な記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、コンピュータ装置による音声認識に関し、より詳細には、周囲環境からの残響がオリジナルの音声に重畳される場合であっても充分に、オリジナル音声を認識するための音声認識装置、音声認識方法、および該制御方法をコンピュータに対して実行させるためのコンピュータ実行可能なプログラムおよび記憶媒体に関する。

【０００２】

【従来技術】

コンピュータ装置による周辺装置の制御性が向上したことにともない、マイクロフォンなどからの音声入力から入力された音声を、自動的に認識するシステムが使用されるようになってきている。上述した音声入力からの音声認識装置は、書類の口述筆記、会議議事録などの書起こし、ロボットとの対話など、外部機械の制御といった種々の用途において利用することができるものと想定することができる。上述した音声認識装置は、本質的には、入力された音声を解析して特徴

量を取得し、取得された特徴量に基づいて音声に対応する単語を選択することにより、音声をコンピュータ装置に対して認識させるものである。音声認識を行う際には、周囲環境からの雑音などの影響を排除するために、種々の方法が提案されている。このための代表的な例としては、ユーザに対してハンド・マイクロフォンまたはヘッドセット型マイクロフォンの使用を義務づけ、収録される音声に重畳される残響やノイズを排除して、入力音声だけを取得する方式を挙げることができる。このような方法では、ユーザが音声収録を行う場合、通常では使用しない余分な機材の使用をユーザに対して要求する。

【０００３】

上述したハンド・マイクロフォンや、ヘッドセット型マイクロフォンの使用をユーザに対して要求する理由としては、発話者がマイクロフォンから離れて発話すると、周囲からの雑音の影響の他にも、周囲環境に応じて生成してしまう残響を挙げることができる。残響がノイズの他に音声信号に重畳されると、音声認識で使用する音声単位の統計モデル：音響モデル(Hidden Markov Model)において、音声認識のミスマッチが生じ、結果的に認識効率の低下を招くことになる。

【０００４】

図９には、音声認識を行う場合に雑音を考慮する代表的な方法を示す。図９に示すように、雑音が存在すると、入力される信号は、音声信号と、音声信号に雑音信号が重畳された出力確率分布を有することになる。多くの場合、雑音は突発的に発生するので、入力信号を取得するためのマイクロフォンと、雑音を取得するためのマイクロフォンとを使用し、いわゆる２チャネルの信号を使用して入力信号から音声信号と、雑音信号とを分離して取得する方法が使用されている。図９に示した従来の音声信号は第１のチャネルにより取得され、雑音信号は、第２のチャネルにより取得されており、２チャネルの信号を使用することによって、雑音のある環境下でも入力された音声信号から、オリジナルの音声信号を認識することが可能とされている。

【０００５】

しかしながら、２チャネル分のデータを使用することにより音声認識装置のハードウエア資源が消費されることに加え、状況によっては２チャネルの入力が可

能でない場合もあるので、常に効率的な認識を可能とするものではない。また都度２チャネルの情報を同時に必要とすることは、現実的な音声認識に対して大きな制限を加えてしまうと言った不都合もある。

【０００６】

従来、音声の伝達経路による影響に対処する方法として、ケプストラム平均減算法(Cepstrum Mean Subtraction: CMS)が使われている。この手法は、例えば電話回線の影響などのように、伝達特性のインパルス応答が比較的短い場合（数msec-数十msec）には有効であるが、部屋の残響のように伝達特性のインパルス応答が長くなった場合（数百msec）には充分な性能が得られないという不都合が知られていた。この理由は、一般的に部屋の残響の伝達特性の長さが、音声認識に用いられる短区間分析の窓幅（10msec-40msec）よりも長くなり、分析区間内で安定したインパルス応答とならないためである。

【０００７】

短区間分析を用いない残響抑制手法としては、複数のマイクロフォンを利用し逆フィルタを設計して音声信号から残響成分を除去する方法も提案されている（M. Miyoshi and Y. Kaneda, “Inverse Filtering of room acoustics,” IEEE Trans. on ASSP, Vol.36, pp.145-152, No.2, 1988）。この方法では、音響伝達特性のインパルス応答が最小位相とならない場合も生じてしまい、現実的な逆フィルタの設計は難しいという不都合がある。また使用環境下において、コストや物理的な配置状況により複数のマイクロフォンを設置できない場合も多い。

【０００８】

また、残響への対応方法は、例えば特開２００２－１５２０９３号公報に開示のエコー・キャンセラのように、種々の方法が提案されている。しかしながら、これらの方法は、音声を２チャネルで入力する必要があり、１チャネルの音声入力で残響に対応することができる方法ではない。さらに、エコー・キャンセラの技術として、特開平９－２６１１３３号公報に記載の方法および装置も知られている。しかしながら、特開平９－２６１１３３号公報において開示される残響処理方法については、同一の残響環境下における複数の場所における音声測定が必要とされる点で、汎用的な方法というわけではない。

【０００９】

また、周囲からのノイズを考慮した音声認識に関しては、例えば共通の出願人に帰属される特許出願、特願２００２－７２４５６号明細書において開示された、フレーム単位で音響モデルを選択することによる、突発性雑音下での音声認識などの方法を使用して対処することも可能である。しかしながら、突発的に発生する雑音ではなく、環境に応じて発生してしまう、残響の特性を有効に利用する音声認識に関して有効な手法は、これまで知られていない。

【００１０】

フレーム内伝達特性Hを予測して、音声認識にフィードバックする方法は、例えば、滝口ら(T. Takiguchi, et. al. “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,” IEEE Trans. on SAP, Vol.9, pp.127-140, No.2, 2001)により報告されている。この方法は、フレーム内における伝達特性Hを使用して残響の影響を反映させ、さらに、音声入力を参照信号としてヘッドセット型のマイクロフォンで入力し、これとは別に残響信号を測定する、２チャネルの測定結果に基づいて、残響を予測する残響予測係数であるαを取得するものである。上述した滝口らの方法を使用することによってもまったく残響の影響を考慮しない場合や、CMS法による処理に比較して充分に高い精度で音声認識を行うことが可能であることが示されているものの、ハンズフリーの環境下で測定された音声信号のみから音声認識を行うことを可能とする方法ではない。

【００１１】

【発明が解決しようとする課題】

しかしながら、手が使用できないユーザや、ヘッドセット型マイクロフォンを携行または着用することができない環境に居るユーザであっても、音声認識を行なうことができれば、音声認識の利用性を大きく広げることができるものと考えられる。また、上述した既存技術はあるものの、既存技術と比較して、さらに音声認識精度を向上させることができれば、音声認識の利用性をさらに拡大することができる。例えば、上述した環境としては、例えば自動車といった車両、航空機などの運転または操縦中や、広い空間内で移動しながら音声認識に基づいて、処理を行う場合、ノート型・パーソナル・コンピュータへの音声入力、キオスク

装置などにおいて離れた位置に配置されたマイクロフォンへの音声入力を行う場合などを挙げることができる。

【００１２】

上述したように、従来の音声認識手法は、少なくともヘッドセット型マイクロフォンやハンド・マイクロフォンなどを使用することが前提とされたものである。しかしながら、コンピュータ装置の小型化や、音声認識の用途が拡大するにつれて、ますます残響を考慮しなければならない環境における音声認識手法が必要とされ、残響が発生する環境においてもハンズフリーでの音声認識機能を可能とする処理がますます要求されて来ている。本発明においては、用語「ハンズフリー」とは、発話者がマイクロフォンの位置に制約を受けず、自由な場所から発話を行うこととして参照する。

【００１３】

【課題を解決するための手段】

本発明は、上述した従来の音声認識の不都合に鑑みてなされたものであり、本発明では、音声認識で使用している音響モデル（Hidden Markov Model）を残響環境下の音声信号に適応させることにより部屋の残響の影響に対処する方法を提案する。本発明では、１つのマイクロフォン（１チャネル）入力で観測された信号を用いて、短区間分析における残響成分の影響を推定する。この方法ではインパルス応答をあらかじめ測定する必要もなく、任意の場所から発話された音声信号のみを用いて、音響モデルを利用した最尤推定により残響成分を推定することを可能とする。

【００１４】

本発明では、本質的に残響や、ノイズの重畳されていない音声信号（以下、本発明では、「フレーム内残響影響を受けた音声モデル」として参照する。）をヘッドセット型のマイクロフォンやハンド・マイクロフォンを使用して実測するのではなく、音声認識で使用している音響モデルを用いて表現し、さらに残響予測係数を尤度最大基準に基づいて推定することによっても、充分な音声認識行うことが可能である、という着想の下になされたものである。

【００１５】

残響が重畳される場合には、入力される音声信号と、音響モデルとは残響の分だけ異なることになる。本発明においては、インパルス応答が長いことを考慮すれば、残響が、過去のフレームにおける音声信号0(ω;tp)に依存しつつ、その時点で判断している音声信号0(ω;t)に重畳されると仮定しても充分に残響をシミュレーションすることができることを見出すことによりなされたものである。本発明においては、残響とは、インパルス応答よりも長時間にわたり音声信号に対して影響を与える信号であり、なおかつ当該残響を与える信号が音声信号を与える話声である、音響的な信号として定義することができる。本発明においてさらに残響を明確に定義することを要するものではないものの、概ね残響は、使用される観測ウィンドウの時間幅との関連で言えば、観測ウィンドウの時間幅よりも長く影響を与える音響的な信号として定義することができる。

【００１６】

ここで、音響モデルとして通常使用される音響モデル・データ(HMMパラメータなど）は、音声コーパスなどを使用して生成される音韻に関連する精度の高い基準信号として捉えることができる。一方で、フレーム内での伝達関数Hは、既存の技術に基づいて充分な精度で予測することができる。本発明では、音響モデルから従来では参照信号として別に入力されていた信号に相当する「フレーム内残響影響を受けた音声モデル」をケプストラムの加算性を使用して生成する。また、残響予測係数αは、選択された音声信号に対して最大尤度を与えるようにして推定することができる。この残響予測係数を使用してユーザに使用する環境に適合した適合音響モデルを生成し、音声認識を実行する。本発明によれば、参照信号としての音声入力を必要とせず、１チャネルからの音声信号のみを使用して音声認識を行うことが可能となる。また、本発明により、発話者がマイクロフォンから離れて発話した場合に問題となる残響の影響に対し、ロバストな音声認識装置および音声認識方法を提供することが可能となる。

【００１７】

すなわち、本発明によれば、コンピュータを含んで構成され音声を認識するための音声認識装置であって、該音声認識装置は、

音声信号から得られる特徴量をフレームごとに格納する記憶領域と、

音響モデル・データおよび言語モデル・データをそれぞれ格納する格納部と、

その時点で処理するべき音声信号よりも前に取得された音声信号から残響音声モデル・データを生成し、残響音声モデル・データを使用して適合音響モデル・データを生成する残響適合モデル生成部と、

前記特徴量と前記適合音響モデル・データと前記言語モデル・データとを参照して音声信号の音声認識結果を与える認識処理手段と

を含む、音声認識装置が提供される。

【００１８】

本発明における前記適合音響モデル生成手段は、ケプストラム音響モデル・データから線形スペクトル音響モデル・データへのモデル・データ領域変換部と、

前記線形スペクトル音響モデル・データに前記残響音声モデル・データを加算して尤度最大を与える残響予測係数を生成する残響予測係数算出部と

を含むことができる。

【００１９】

本発明では、残響音声モデル・データを生成する加算部を含み、前記加算部は、前記音響モデルのケプストラム音響モデル・データおよびフレーム内伝達特性のケプストラム音響モデル・データを加算して「フレーム内残響影響を受けた音声モデル」を生成することができる。

【００２０】

本発明における前記加算部は、生成された「フレーム内残響影響を受けた音声モデル」を前記モデル・データ領域変換部へと入力し、前記モデル・データ領域変換部に対して「フレーム内残響影響を受けた音声モデル」の線形スペクトル音響モデル・データを生成させることができる。

【００２１】

本発明における前記残響予測係数算出部は、入力された音声信号から得られた少なくとも１つの音韻と、前記残響音声モデル・データとを使用して線形スペクトル音響モデル・データに基づいて残響予測係数の尤度を最大化させることができる。本発明における前記音声認識装置は、隠れマルコフ・モデルを使用して音声認識を実行することが好ましい。

【００２２】

本発明によれば、コンピュータを含んで構成され音声を認識するための音声認識装置に対して音声認識を実行させるための方法であって、前記方法は、前記音声認識装置に対して、

音声信号から得られる特徴量をフレームごとに記憶領域に格納させるステップと、

その時点で処理するべき音声信号よりも前に取得された音声信号を前記格納部から読み出して残響音声モデル・データを生成し、格納部に格納された音響モデル・データを処理して適合音響モデル・データを生成して記憶領域に格納させるステップと、

前記特徴量と前記適合音響モデル・データと格納部に格納された言語モデル・データとを読み込んで音声信号の音声認識結果を生成させるステップと

を含む、音声認識方法が提供される。

【００２３】

本発明によれば、前記適合音響モデル・データを生成するステップは、加算部により前記読み出された音声信号とフレーム内伝達特性値との合計値を算出するステップと、

前記加算部により算出された前記合計値をモデル・データ領域変換部に読み込ませ、ケプストラム音響モデル・データから線形スペクトル音響モデル・データへと変換させるステップと、を含むことができる。

【００２４】

本発明においては、加算部に対して前記線形スペクトル音響モデル・データと前記残響音声モデル・データとを読み込ませ加算して、尤度最大を与える残響予測係数を生成させるステップと、を含むことができる。本発明においては、前記線形スペクトル音響モデル・データへと変換させるステップは、前記加算部に対して、前記音響モデル・データのケプストラム音響モデル・データおよびフレーム内伝達特性のケプストラム音響モデル・データを加算して「フレーム内残響影響を受けた音声モデル」を生成するステップを含むことができる。

【００２５】

本発明における前記残響予測係数を生成させるステップは、前記加算部により生成された前記「フレーム内残響影響を受けた音声モデル」の線形スペクトル音響モデル・データと前記残響音声モデル・データとの合計値が音声信号から生成され格納された少なくとも１つの音韻に対して最大の尤度を与えるように残響予測係数を決定するステップを含むことができる。

【００２６】

本発明においては、上記の音声認識方法をコンピュータに対して実行させるためのコンピュータ可読なプログラムおよびコンピュータ可読なプログラムを記憶した、コンピュータ可読な記憶媒体が提供される。

【００２７】

【発明の実施の形態】

以下、本発明を図面に示した実施の形態をもって説明するが、本発明は、後述する実施の形態に限定されるものではない。

【００２８】

Ａ：隠れマルコフ・モデルを使用する音声認識の概説

図１には、本発明において使用する、隠れマルコフ・モデル(Hidden Markov Model:HMM)を使用した音声認識を概略的に説明する。音響モデルは、単語または文が、音韻(phoneme)の連続として構築されており、それぞれの音韻に対して、典型的には３状態を付与し、これらの状態間の遷移確率を規定することにより、音韻の連続する単語または文を検索するオートマトンとして考えることができる。図１に示した実施の形態は、説明のために３つの音韻S1～S3が示されており、状態S1から状態S2への遷移確率Pr(S1|S0)は、0.5であり、また、状態S2から状態S3への遷移確率Pr(S3|S2)は、0.3であるものとして示されている。

【００２９】

それぞれの状態S1～S3には、例えば混合ガウス分布により与えられる音韻に関連して決定される出力確率が割り当てられており、図１に示した実施の形態では、状態S1から状態S3には、k1～k3の混合要素が使用されているのが示されている。また、図１には、k1～k3で示される状態S1に対応する混合ガウス分布の出力確率分布が示されている。それぞれの混合要素には、重みw1～w3が与えられており

、特定の話者に対して適切に適応させることができるようにされている。上述した音響モデルを使用すると、出力確率は、音声信号をアルファベットの「0」で表し、HMMパラメータのセットをλで表すと、Pr(0|λ)で与えられるものとして定義される。

【００３０】

図２には、本発明における出力確率テーブルを生成するための処理を示す。図２に示した実施の形態では、例えば状態S1から状態S3までに至る出力確率は、音声信号から得られる特徴量系列｛α　β　α｝を使用して、図２のようなトレリスを構成させ、ビタビ・アルゴリズム、フォワード・アルゴリズム、ビームサーチ・アルゴリズムなどを使用して算出することができる。より一般的には、所定のフレームtでの音声信号0t、状態s、およびHMMパラメータのセットλとすれば、音声信号に対して各状態に基づく出力確率は、出力確率テーブルとして与えられることになる。

【００３１】

【数１】

$$\Pr(O|\lambda)=\sum_{allS}\prod_{t=1}^{T}\Pr(o_t|s_t,s_{t-1},\lambda)\Pr(s_t|s_{t-1},\lambda) \qquad (1)$$

HMMによる音声認識では、上述した出力確率テーブルを使用して、最尤の音韻列を検索することにより、出力結果である単語または文を決定する。それぞれの状態は、混合ガウス分布で記述されるものの、最初の音韻から最後の音韻までの間は、状態遷移確率による尤度によって決定されることになる。なお、一般的なHMMによる音声認識については、例えば鹿野ら、「音声・音情報のディジタル信号処理」、昭晃堂、ISBN4-7856-2014を参照することができる。

【００３２】

Ｂ：本発明の音声認識方法における処理

図３には、本発明の音声認識方法の概略的な手順を示したフローチャートを示す。図３に示されるように、本発明の音声認識方法の処理は、ステップＳ１０において音声信号の入力を受け取り、ステップＳ１２において、音響モデル・データとフレーム内伝達特性とから残響のない場合の「フレーム内残響影響を受けた

音声モデル」を生成する。ステップＳ１４では、残響予測係数αと、過去の音声信号とを使用して残響音声モデル・データを生成する（α×0|ω;tp|）。

【００３３】

生成された残響音声モデル・データは、ステップＳ１６においてステップＳ１２で与えられた「フレーム内残響影響を受けた音声モデル」と線形スペクトル音響モデル・データとして加算された後、音声信号を処理して得られた選択された単語または文の音韻との間における最尤値が得られるように残響予測係数αを決定する。ステップＳ１８では、決定された残響予測係数αおよび過去のフレームの音声信号0(ω;tp)とを使用して、残響の絶対値を取得し、フレーム内残響影響を受けた音声モデルの平均値ベクトルμに加算して、$\mu' = \mu + \alpha \times 0(\omega;tp)$を計算し、フレーム外の残響影響成分も含む音声モデルを生成させ、他のパラメータとセットとして格納させる。その後、ステップＳ２０において、音声信号と、適合音響モデル・データとを使用して音声認識を実行させ、ステップＳ２２において認識結果を出力させる。

【００３４】

図４は、本発明の図３において説明した処理の概略的な処理を示した図である。まず、音響モデル・データおよびフレーム内伝達特性のケプストラムを加算して、「フレーム内残響影響を受けた音声モデル」のデータ（以下、本発明では、「フレーム内残響影響を受けた音声モデル」として参照する。）を作成する。生成された音声モデル・データに対して、離散フーリエ変換といった方法および指数化処理を施して線形スペクトル音響モデル・データに変換する。さらに、残響予測係数αは、変換後のスペクトル・データにおいて選択された音声信号に含まれる音韻の特徴量に対して尤度を最大とするように決定される。この際の設定としては、種々の方法を使用することができるものの、例えば、一定の単語や、一定の文を使用して、適宜決定することができる。決定された残響予測係数αは、元々音声認識装置が格納していた音響モデル・データとともに適合音響モデル・データを作成するために使用され、生成された線形スペクトル領域での音響モデル・データが対数変換および逆フーリエ変換を行うことによりケプストラムとされ、音声認識を実行させるために格納される。

【００３５】

ここで、音声信号が残響を含む音声である場合について考える。残響が音声に重畳される場合に、その時点で観測される、周波数ω、フレーム番号tの音声信号O'（ω;t)は、過去のフレームの音声信号O(ω;tp)を使用して、下記式（２）で示されることが知られている（「中村、滝口、鹿野、「短区間スペクトル分析における残響補正に関する検討」、日本音響学会講演論文集、平成10年3月、3-6-11」）。

【００３６】

【数２】

$$O'(\omega;t) \cong S(\omega;t)\cdot H(\omega)+\alpha\cdot O(\omega;t-1) = \exp[\cos\{S_{cep}(c;t)+H_{cep}(c)\}]+\alpha\cdot O(\omega;t-1) \qquad (2)$$

上記式中、Sは、本発明においては音声コーパスなどを使用して生成された標準的な音響モデルを使用することができ、これを本発明においてクリーン音声信号として参照する。Hは、同一フレーム内での伝達特性の予測値を使用する。また、αは、過去のフレームからその時点で評価するフレームへと重畳されることになる残響の割合を示す残響予測係数である。添え字のcepは、ケプストラムを意味している。

【００３７】

従来では、本発明では、音声認識で使用している音響モデル・データを参照信号の代わりに使用する。さらにフレーム内伝達特性Hを予測値として取得し、残響予測係数を尤度最大基準に基づいて選択された音声信号を使用して決定することにより、適合音響モデル・データを生成する。

【００３８】

残響が重畳される場合には、入力音声信号と、音響モデル・データとは残響の分だけ異なることになる。本発明においては、インパルス応答が長いことを考慮すれば、残響が、直前のフレームにおける音声信号O(ω;tp)に依存しつつ、その時点で判断している音声信号O(ω;t)に重畳されると仮定しても充分に残響をシミュレーションすることができることに着目した。すなわち上記式（２）を使用して、音声信号に対して所定の音響モデル・データとαとの値から尤度が最も高

くなる音響モデル・データを決定することにより、対応する言語モデル・データを使用して、１チャネルからの音声信号のみを使用して音声認識を行うことが可能となる。

【００３９】

また、音響モデル・データに対してフレーム内伝達特性Hの加算は、スペクトル領域では、コンボリューションにより得られるものの、ケプストラム領域に変換すれば加算条件が成立するので、他の方法によりフレーム内伝達特性Hの推定ができれば、容易に音響モデル・データとの加算性を使用して、容易かつ精度良くすでに登録されている音響モデル・データのケプストラム領域のデータとの加算によりフレーム内伝達特性Hを考慮した音響モデル・データを決定できる。

【００４０】

以下、クリーン音声信号SのHMMにおけるパラメータの集合を$\lambda(s),cep$、フレーム内伝達特性HのHMMパラメータの集合を$\lambda(h'),cep$、適応後の音響モデル・データのHMMパラメータの集合を$\lambda(0),cep$とする。本発明においては、音響モデル・データのうち、出力確率分布のみに注目するので、所定のHMMの状態jのk番目の出力確率分布の平均値を$\mu_{j,k}$、分散を$\sigma^2(S)_{j,k}$、重みを$w_{j,k}$とした場合、$\lambda(s)$を、$\lambda(s)=\{\mu_{j,k}、\sigma^2(S)_{j,k}、w_{j,k}\}$で表わすものとする。通常、これらの音響モデル・データのHMMパラメータは、音声認識に最もよく適しているケプストラムとされて、音声認識に適用される。

【００４１】

図３のステップＳ１２におけるフレーム内伝達特性の推定は、例えば、本発明における特定の実施の形態では、T. Takiguchi, et. al. “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,”、IEEE Trans. on SAP, Vol.9, No.2, 2001に記載された方法において、便宜的に残響が存在しないものとして$\alpha=0$と設定して得られたフレーム内伝達関数Hを使用することができる。生成されたフレーム内伝達関数Hは、離散フーリエ変換(Discrete Fourie Transformation)および指数化処理を行って、ケプストラム領域に変換して、後述する記憶領域に適時的に格納させておくことができる。

【００４２】

また、残響予測係数αを、尤度に基づいて算出する場合には、種々の方法を使用することができる。本発明において説明している特定の実施の形態では、EMアルゴリズム（“An inequality and associated maximization technique in statistical estimation of probabilistic function of a Markov process”, Inequalities, Vol. 3, pp. 1-8, 1972）を使用し、最大尤度の予測値であるα’を算出することができる。

EMアルゴリズムを使用する残響予測係数αの計算処理は、EMアルゴリズムのE-ステップと、M-ステップとを使用して実行される。まず、本発明においては、線形スペクトル領域に変換されたHMMパラメータのセットを使用して、E-ステップにおいて下記式（３）で示されるQ関数を計算する。

【００４３】

【数３】

$$Q(\alpha'|\alpha) = E[\log \Pr(O,s,k|\lambda_{(SH),lin},\alpha')|\lambda_{(SH),lin},\alpha]$$

$$=\sum_{p}\sum_{n}\sum_{s_{p,n}}\sum_{m_{p,n}}\frac{\Pr(O_{p,n},s_{p,n},m_{p,n}|\lambda_{(SH),lin},\alpha)}{\Pr(O_{p,n}|\lambda_{(SH),lin},\alpha)}\cdot\log\Pr(O_{p,n},s_{p,n},m_{p,n}|\lambda_{(SH),lin},\alpha') \quad (3)$$

上記式中、pは、HMMパラメータのインデックス（例えば所定の音韻などを表す。）であり、$O_{p,n}$は、音韻pに関連するn番目の観測系列とする。また$s_{p,n}$、$m_{p,n}$は、$O_{p,n}$それぞれに対する状態系列および混合要素の系列とする。$\lambda_{(SH),lin}$の音韻pの状態jのk番目の出力確率分布（混合ガウス分布）の平均値、分散、重みを下記式（４）とし、

【００４４】

【数４】

$$\{\mu_{(SH),p,j,k},\sigma^2_{(SH),p,j,k},w_{(SH),p,j,k}\} \quad (4)$$

各々の次元数をDとした場合、上記Q関数の出力確率分布のみに関する項に注目すると、Q関数は、下記式（５）で示される。

【００４５】

【数５】

$$Q(\alpha'|\alpha) = -\sum_{p}\sum_{n}\sum_{j}\sum_{k}\sum_{t}\gamma_{p,n,j,k,t}\left\{\frac{1}{2}\log(2\pi)^{D}\sigma^{2}_{(SH),p,j,k} + \frac{\{O_{p,n}(t)-\mu_{(SH),p,j,k}-\alpha'\cdot O_{p,n}(t-1)\}^{T}\{O_{p,n}(t)-\mu_{(SH),p,j,k}-\alpha'\cdot O_{p,n}(t-1)\}}{2\sigma^{2}_{(SH),p,j,k}}\right\} \quad (5)$$

上記式中、tは、フレーム番号を表す。またγ p, n, j, k, tは、下記式（６）で与えられる確率である。

【００４６】

【数６】

$$\gamma_{p,n,j,k,t} = \Pr(O_{p,n}(t), j, k \mid \lambda_{(SH),lin}, \alpha) \quad (6)$$

次に、EMアルゴリズムにおけるM-step (Maximization)で、Q関数をα’に関して最大にする。

【００４７】

【数７】

$$\alpha' = \arg\max_{\alpha'} Q(\alpha'|\alpha) \quad (7)$$

最大尤度のα’は、得られたQを、α’で偏微分して、極大値を求めることにより得ることができる。その結果、α’は、下記式（８）で与えられる。

【００４８】

【数８】

$$\alpha' = \frac{\sum_{p}\sum_{n}\sum_{j}\sum_{k}\sum_{t}\gamma_{p,n,j,k,t}\frac{O_{p,n}(t)\cdot O_{p,n}(t-1)-O_{p,n}(t-1)\cdot\mu_{(SH),p,j,k}}{\sigma^{2}_{(SH),p,j,k}}}{\sum_{p}\sum_{n}\sum_{j}\sum_{k}\sum_{t}\gamma_{p,n,j,k,t}\frac{O^{2}_{p,n}(t-1)}{\sigma^{2}_{(SH),p,j,k}}} \quad (8)$$

本発明においては、音韻pごとにα’を推定することもでき、この場合には、下記式（９）で与えられるように、音韻pでの総和を算出する前の値を使用することで音韻ごとのα’を取得することもできる。

【００４９】

【数９】

$$\alpha_p' = \frac{\sum_n \sum_j \sum_k \sum_t \gamma_{p,n,j,k,t} \dfrac{O_{p,n}(t)\cdot O_{p,n}(t-1) - O_{p,n}(t-1)\cdot \mu_{(SH),p,j,k}}{\sigma^2_{(SH),p,j,k}}}{\sum_n \sum_j \sum_k \sum_t \gamma_{p,n,j,k,t} \dfrac{O^2_{p,n}(t-1)}{\sigma^2_{(SH),p,j,k}}} \qquad (9)$$

いずれの残響予測係数を使用するかについては、認識の効率や認識速度といった特定の装置および要求に応じて決定することができる。また、HMM状態ごとにα’を求めることも式（８）、式（９）と同様に可能である。上述した計算処理を実行させることにより、オリジナルの音響モデルのパラメータのみを使用して、発話者から離れた1チャネル入力の音声信号0(t)のみから、残響予測係数αを得ることができる。

【００５０】

Ｃ：本発明の音声認識装置とその処理方法

図５には、本発明の音声認識装置の概略的なブロック図を示す。本発明の音声認識装置１０は、概ね中央処理装置(CPU)を含むコンピュータを使用して構成されている。図５に示すように、本発明の音声認識装置１０は、音声信号取得部１２と、特徴量抽出部１４と、認識処理部１６と、適合音響モデル・データ生成部１８とを含んで構成されている。音声信号取得部１２は、図示しないマイクロフォンといった入力手段から入力される音声信号をA/Dコンバータなどによりディジタル信号とし、振幅を時間フレームと対応づけて適切な記憶領域２０に格納させている。特徴量抽出部１４は、モデル・データ領域変換部２２を含んで構成されている。

【００５１】

モデル・データ領域変換部２２は、図示しないフーリエ変換手段と、指数化手段と、逆フーリエ変換手段とを含んで構成されており、記憶領域２０に格納された音声信号を読み出して、音声信号のケプストラムを生成させ、記憶領域２０の適切な領域に格納する。また、特徴量抽出部１４は、生成された音声信号のケプストラムから特徴量系列を取得し、フレームに対応させて格納する。

【００５２】

図５に示した本発明の音声認識装置１０は、さらに、音声コーパスなどを使用して生成された、HMMに基づく音響モデル・データを格納する音響モデル・データ格納部２４と、テキスト・コーパスなどから得られた言語モデル・データを格納する言語モデル・データ格納部２６と、本発明により生成された適合音響モデル・データを格納する、適合音響モデル・データ生成部１８とを含んで構成されている。

認識処理部１６は、本発明においては、適合音響モデル・データを適合音響モデル・データ格納部２８から読み出し、言語モデル・データを言語モデル・データ格納部２６から読み出し、読み出された各データを、音声信号のケプストラムに基づき、尤度最大化を使用して音声認識を実行することができる構成とされている。

【００５３】

本発明において使用することができる音響モデル・データ格納部２４と、言語モデル・データ格納部２６と、適合音響モデル・データ格納部２８とは、それぞれハードディスクといった記憶装置に構築されたデータベースとすることができる。また、図５に示された適合音響モデル・データ生成部１８は、本発明における上述の処理により適合音響モデル・データを作成して、適合音響モデル・データ格納部２８へと格納させている。

【００５４】

図６は、本発明において使用される適合音響モデル・データ生成部１８の詳細な構成を示した図である。図６に示すように、本発明において使用する適合音響モデル・データ生成部１８は、バッファ・メモリ３０と、モデル・データ領域変換部３２ａ、３２ｂと、残響予測係数算出部３４と、加算部３６ａ、３６ｂと、生成部３８とを含んで構成されている。適合音響モデル・データ生成部１８は、その時点で処理を行うフレームｔよりも過去の所定の観測データを読み込んで、残響予測係数αを乗じてバッファ・メモリ３０に一旦格納させる。同時に、音響モデル・データ格納部２４から音響モデル・データを読み込み、予め計算しておいたフレーム内伝達特性Hのケプストラム音響モデル・データを、記憶領域２０からバッファ・メモリ３０へと読み込む。

【００５５】

バッファ・メモリ３０に格納された音響モデル・データと、フレーム内伝達特性のデータは、いずれもケプストラム音響モデル・データとされているので、これらのデータは、加算部３６ａへと読み込まれ、加算が実行され、「フレーム内残響影響を受けた音声モデル」が生成される。「フレーム内残響影響を受けた音声モデル」は、モデル・データ領域変換部３２ａへと送られ、線形スペクトル音響モデル・データに変換された後、加算部３６ｂへと送られる。加算部３６ｂは、さらに過去の観測データに残響予測係数を乗じたデータを読み込んで、「フレーム内残響影響を受けた音声モデル」の線形スペクトル音響モデル・データと加算を実行する。

【００５６】

加算部３６ｂにおいて生成された加算データは、予め選択された音韻などに対応する音響モデル・データを格納した残響予測係数算出部３４へと送られ、EMアルゴリズムを使用して尤度最大となるように、残響予測係数αを決定する。決定された残響予測係数αは、線形スペクトル音響モデル・データに変換または線形スペクトルのまま格納された音響モデル・データと共に、生成部３８へと渡され、適合音響モデル・データとして作成される。作成された適合音響モデル・データは、モデル・データ領域変換部３２ｂへと送られ、線形スペクトル音響モデル・データからケプストラム音響モデル・データへと変換された後、適合音響モデル・データ格納部２８へと格納される。

【００５７】

図７は、本発明の音声認識装置により実行される音声認識方法の処理を示す概略的なフローチャートである。図７に示すように、本発明の音声認識装置が実行する認識処理は、ステップＳ３０において、残響の重畳された音声信号をフレームごとに取得して、少なくともその時点で処理を実行させる処理フレームと、それ以前のフレームとを、適切な記憶領域に格納させる。ステップＳ３２において音声信号から特徴量を抽出し、音響モデル・データおよび言語モデル・データによる音声信号の検索のために使用するデータを取得して、ケプストラム音響モデル・データとして適切な記憶領域に格納する。

【００５８】

一方、ステップＳ３４は、ステップＳ３２と並列的に処理を行うことができ、過去のフレームの音声信号および音響モデル・データを適切な記憶領域から読み出し、ケプストラム領域への変換処理および線形スペクトル領域への変換処理を使用して、適合音響モデル・データを作成し、適切な記憶領域へと予め格納しておく。ステップＳ３６において、適合音響モデル・データと、音声信号から得られた特徴量とを使用して最大尤度を与える音韻を決定し、ステップＳ３８において決定された音韻に基づいて言語モデル・データを使用して、認識結果を生成し、適切な記憶領域に格納させる。同時に、その時点での尤度の合計を格納する。その後、ステップＳ４０において、処理するべきフレームが残されているかを判断し、処理するべきフレームがない場合(no)には、ステップＳ４２において尤度の和が最大の単語または文を認識結果として出力する。また、ステップＳ４０の判断において処理するべきフレームが残されている場合(yes)には、ステップＳ４４において、残されているフレームの観測データを読み込んで、特徴量を抽出し、ステップＳ３６へと処理を戻し、処理を繰り返すことにより、単語または文の認識を完了させる。

【００５９】

図８には、本発明の音声認識装置を、ノート型パーソナル・コンピュータ４０として構成させた実施の形態を示す。ノート型パーソナル・コンピュータ４０には、表示部上側に内蔵マイクロフォン４２が配設されており、ユーザからの音声入力を受け取ることができる構成とされている。ユーザは、例えばオフィスや自宅などに設置されたマウスまたはタッチパッドといったポインタ手段４４を使用して表示部に表示されたカーソルを移動させ種々の処理を実行させる。

【００６０】

ここで、ユーザは、音声認識を使用する、例えばＩＢＭ社製のソフトウエア(ViaVoice：登録商標）を使用したワードプロセッサ・ソフトウエアにより、ディクテーションを行うことを希望するものとする。このときユーザが、例えばアプリケーションを起動するためのアプリケーション・アイコン４６にマウス・カーソルを重ね合わせ、マウス４４をクリックすると、ワードプロセッサ・ソフトウ

エアは、ViaVoiceソフトウエアと同時に起動される。本発明の特定の実施の形態では、ViaVoiceソフトウエアに対して本発明の音声認識プログラムがモジュールとして搭載されている。

【００６１】

従来では、ユーザは、ヘッドセット型マイクロフォンや、ハンド・マイクロフォンを使用して、残響や周囲ノイズの影響を避けながら音声入力する。また、ユーザは、周囲ノイズや残響と入力音声とを別々に入力して、音声入力を行うことが要求されることになる。しかしながら、本発明の図８に示されたノート型パーソナル・コンピュータ４０を使用した音声認識方法では、ユーザは、本発明にしたがい、内蔵マイクロフォン４２により入力を行うだけで、音声認識を介したディクテーションを行うことが可能となる。

【００６２】

図８は、本発明をノート型パーソナル・コンピュータに対して適用した実施の形態を示しているものの、本発明は、図８に示した以外にも、区画された比較的狭い部屋の中で音声対話式に処理を進めるためのキオスク装置や、乗用車、航空機などにおけるディクテーションや、コマンド認識など、周囲ノイズの常態的な重畳よりも残響の影響が大きな、比較的狭い空間内における音声対話型処理に適用することができる。また、本発明の音声認識装置は、ネットワークを介して、非音声処理を行う他のサーバ・コンピュータまたは音声処理対応型のサーバ・コンピュータとの通信を行うことも可能である。上述したネットワークとしては、ローカル・エリア・ネットワーク(LAN)、ワイド・エリア・ネットワーク(WAN)、光通信、ISDN、ADSLといった通信インフラ基盤を使用したインターネットなどを挙げることができる。

【００６３】

本発明の音声認識方法では、時系列的に連続して入力される音声信号を使用するのみで、マイクロフォンを複数使用して別に参照信号を格納し、処理するための余分な処理ステップおよびそのためのハードウエア資源を必要としない。また、参照信号を「フレーム内残響影響を受けた音声モデル」として取得するためのヘッドセット型マイクロフォンやハンド・マイクロフォンを使用することなく、

音声認識の利用性を拡大することを可能とする。

【００６４】

これまで、本発明の図面に示した特定の実施の形態に基づいて説明してきたが、本発明は、説明した特定の実施の形態に限定されるものではなく、各機能部または機能手段は、コンピュータに対してプログラムを実行させることにより実現されるものであり、図面に示した機能ブロックごとの構成として必ずしも構成されなければならないものではない。また、本発明の音声認識装置を構成させるためのコンピュータ可読なプログラミング言語としては、アセンブラ語、FORTRAN、C言語、C++言語、Java（登録商標）などを挙げることができる。また、本発明の音声認識方法を実行させるためのコンピュータ実行可能なプログラムは、ROM、EEPROM、フラッシュ・メモリ、CD-ROM、DVD、フレキシブル・ディスク、ハードディスクなどに格納して頒布することができる。

【００６５】

Ｄ：実施例

以下、本発明を具体的な実施例を使用して説明する。残響下での音声を作成するために、実際に部屋で測定したインパルス応答を使用した。実施例、参考例および比較例共に、残響時間としては300msecに対応するフレームの値を用いた。音源位置は、マイクからの距離を2mとし、正面方向からマイクロフォンに向かって話声を入力させた。信号分析条件は、サンプリング周波数12kHz、ウィンドウ幅32msec、分析周期8msecを使用した。音響特徴量としては、16次元のMFCC(Mel Frequency Cepstral Coefficient)を用いた。

【００６６】

分析周期を8msecとしたので、ウィンドウ間で重なりが生じないように、4フレーム分ずらした過去の音声信号を、残響信号の処理のために使用した。実施例、参考例および比較例ともに、使用した入力音声信号は、55個の音韻から生成させた。また、残響予測係数αの算出は、入力した音声入力信号のうち、一単語分の音韻を使用して尤度最大を計算させ、得られた残響予測係数αを、すべての音声認識について適用した。以下に、500単語を認識させた場合の、認識成功率の結果を示す。

【００６７】

【表１】

	実施例	参考例	比較例１	比較例２
手法	本発明	滝口ら	CMS	残響補正なし
認識成功率	92.8%	91.2%	86.0%	54.8%

上記表１に示されるように、残響補正なしの場合（比較例２）では、54.8%の結果が得られた。一方で、本発明（実施例）によれば、認識成功率は、92.8%まで高めることができた。この結果は、滝口らの参考例(前掲：T. Takiguchi, et. al. “HMM-Separation-Based Speech Recognition for a Distant Moving Speaker,” IEEE Trans. on SAP, Vol.9, pp.127-140, No.2, 2001)により得られた、参照信号を使用する2チャネル・データを使用する場合よりも僅かに良好な結果が得られている。また比較例１としてCMS法（ケプストラム平均減算法）を使った場合では、認識成功率が86%と、本発明の実施例よりも低い結果が得られた。すなわち、本発明によれば、１チャネル・データを使用するにもかかわらず、従来よりも良好な認識成功率を提供できることが示された。

【図面の簡単な説明】

【図１】　隠れマルコフ・モデル(Hidden Markov Model:HMM)を使用した音声認識を概略的に説明した図。

【図２】　音声信号に対して各状態に基づく出力確率テーブルを形成するための処理を概略的に説明した図。

【図３】　本発明の音声認識方法の概略的な手順を示したフローチャート。

【図４】　図３において説明した処理の概略的な処理を示した図。

【図５】　本発明の音声認識装置の概略的なブロック図。

【図６】　本発明において使用される適合音響モデル・データ生成部の詳細な構成を示した図。

【図７】　本発明の音声認識装置により実行される音声認識方法の処理を示す概略的なフローチャート。

【図８】　本発明の音声認識装置を、ノート型パーソナル・コンピュータとして構成させた実施の形態を示した図。

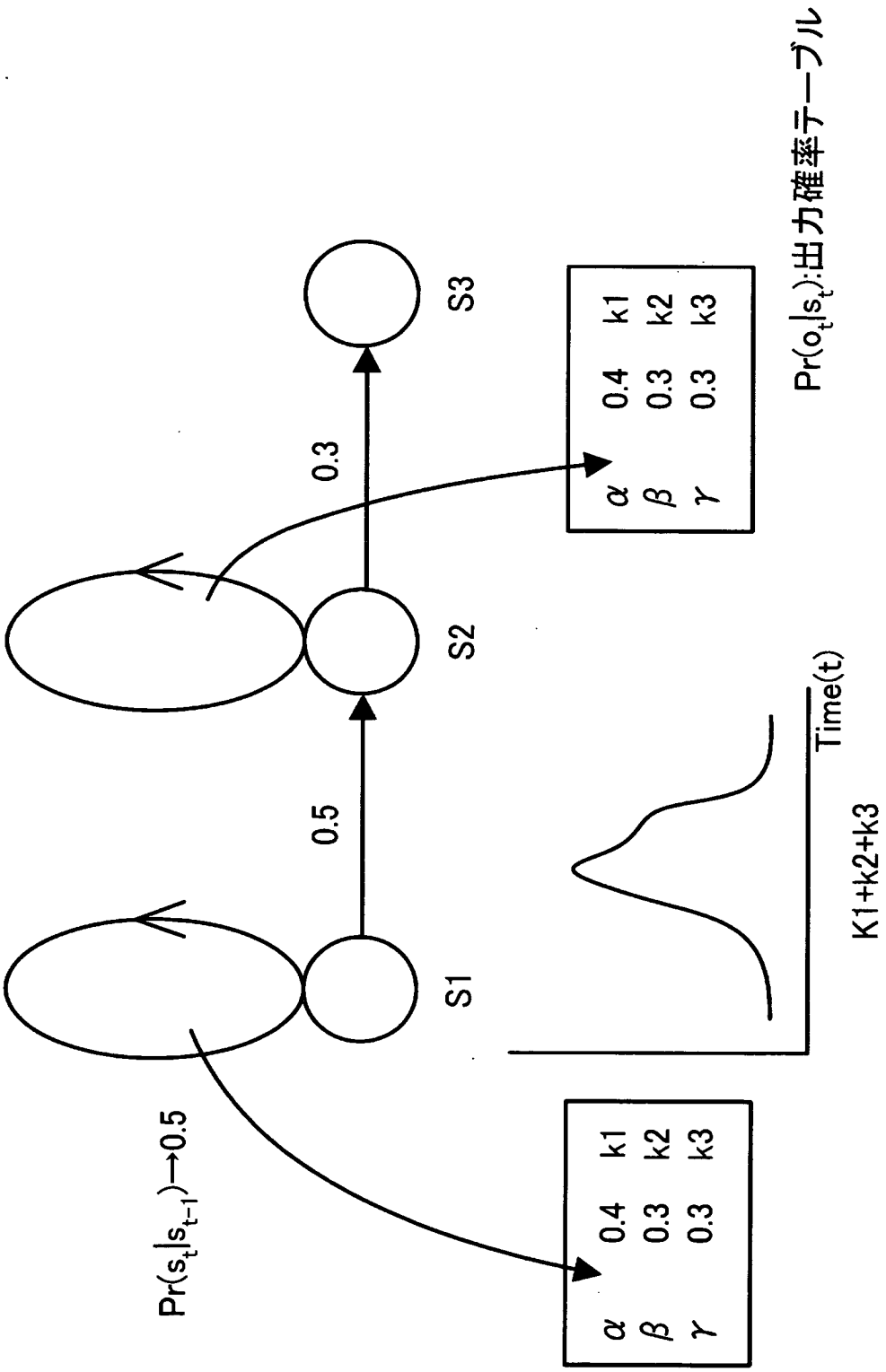
【図９】　音声認識を行う場合に雑音を考慮する代表的な方法を示した図。

【符号の説明】

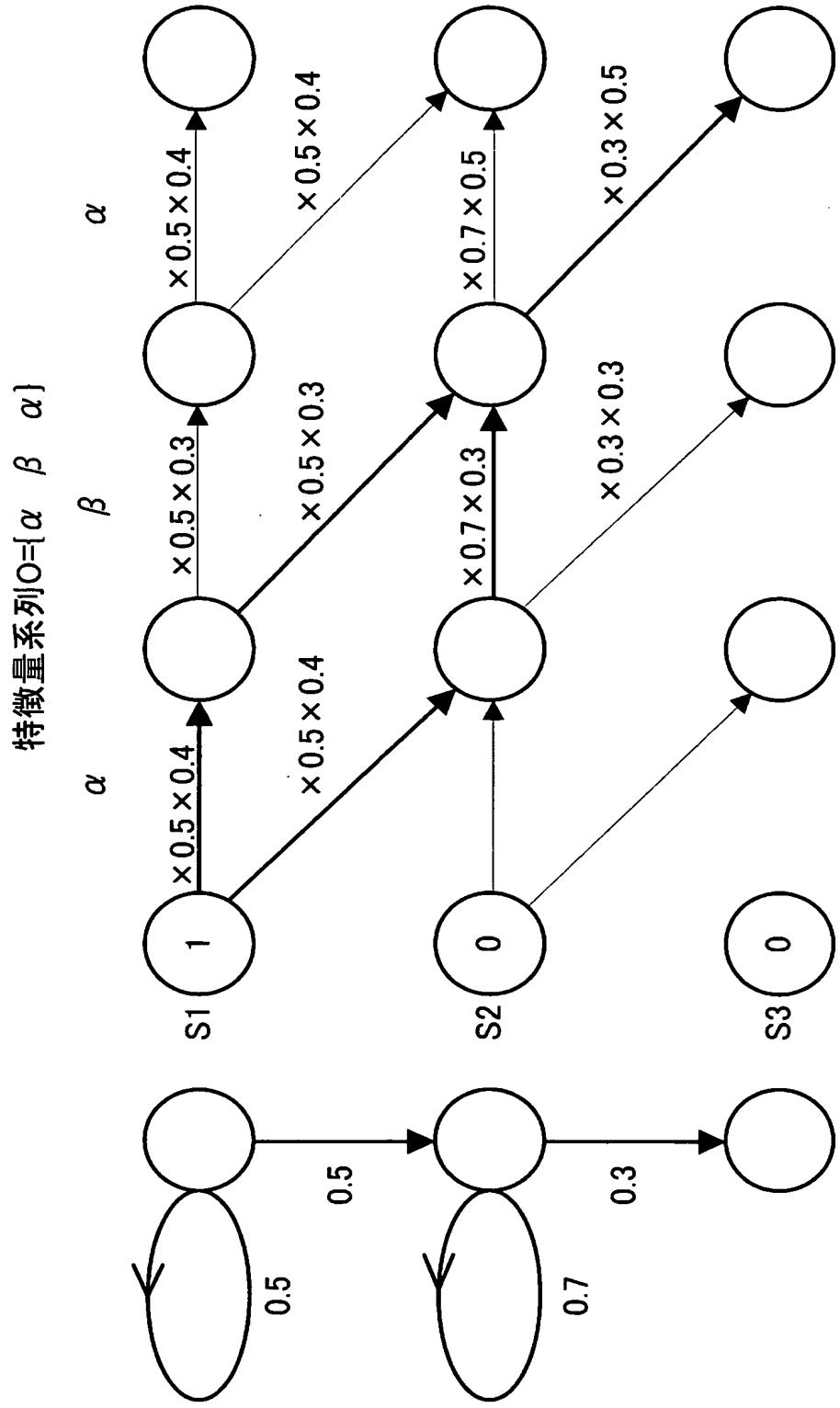
１０…音声認識装置、１２…音声信号取得部、１４…特徴量抽出部、１６…認識処理部、１８…適合音響モデル・データ生成部、２０…記憶領域、２２…モデル・データ領域変換部、２４…音響モデル・データ格納部、２６…言語モデル・データ格納部、２８…適合音響モデル・データ格納部、３０…バッファ・メモリ、３２…モデル・データ領域変換部、３４…残響予測係数算出部、３６…加算部、３８…生成部、４０…ノート型パーソナル・コンピュータ、４２…内蔵マイクロフォン、４４…ポインタ手段、４６…アプリケーション・アイコン

【書類名】　図面

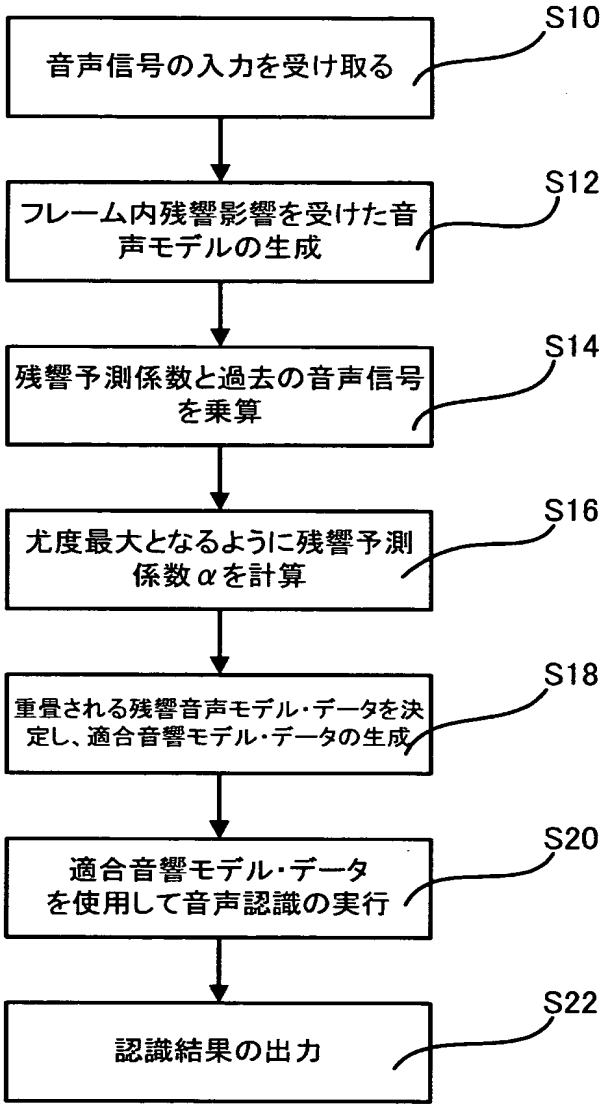
【図１】


S1
S2
S3
0.5
0.3
Pr(s_t|s_t-1)→0.5
α β γ
0.4 0.3 0.3
k1 k2 k3
Pr(o_t|s_t):出力確率テーブル
Time(t)
K1+k2+k3

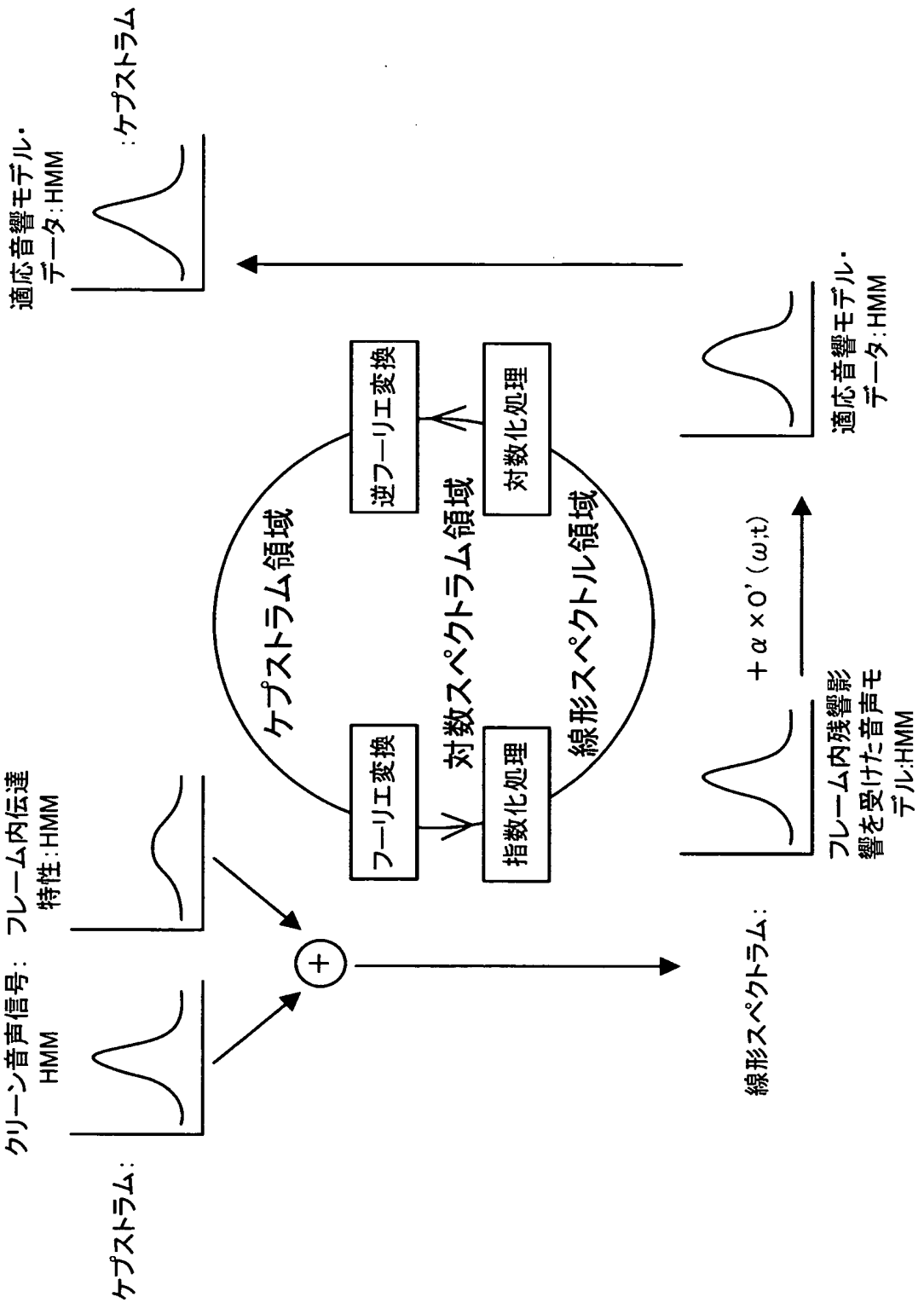
【図２】


特徴量系列O={α β α}
α
β
α
×0.5×0.4
×0.5×0.3
×0.5×0.4
×0.5×0.4
×0.5×0.3
×0.5×0.4
×0.7×0.5
×0.7×0.3
×0.3×0.5
×0.3×0.3
1
0
0
S1
S2
S3
0.5
0.5
0.3
0.5
0.7

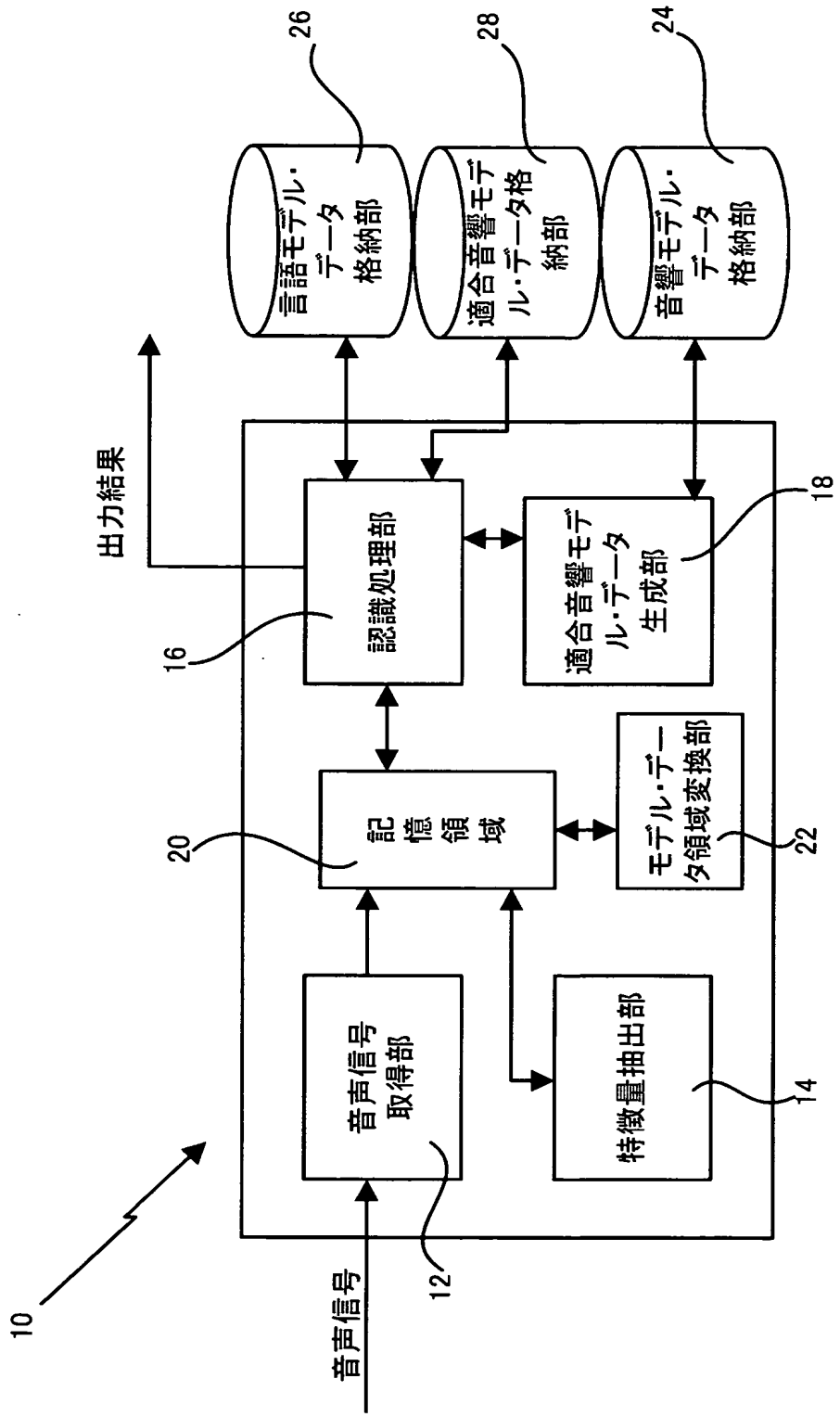
【図３】


S10
音声信号の入力を受け取る
S12
フレーム内残響影響を受けた音声モデルの生成
S14
残響予測係数と過去の音声信号を乗算
S16
尤度最大となるように残響予測係数αを計算
S18
重畳される残響音声モデル・データを決定し、適合音響モデル・データの生成
S20
適合音響モデル・データを使用して音声認識の実行
S22
認識結果の出力

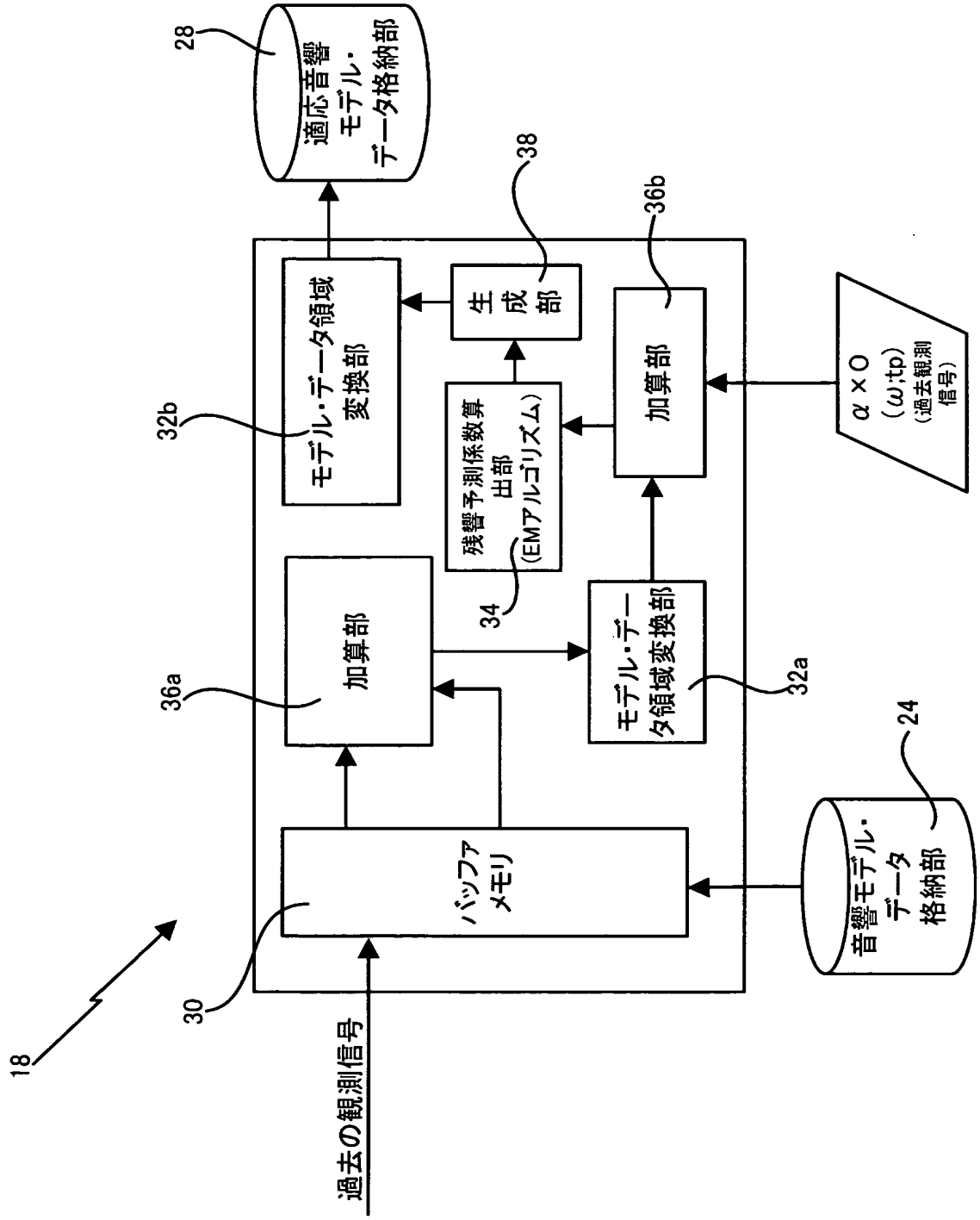
【図４】


適応音響モデル・
データ:HMM
:ケプストラム
クリーン音声信号:
HMM
フレーム内伝達
特性:HMM
ケプストラム:
ケプストラム領域
逆フーリエ変換
対数化処理
フーリエ変換
指数化処理
対数スペクトラム領域
線形スペクトル領域
＋α×O'(ω;t)
線形スペクトラム:
フレーム内残響影
響を受けた音声モ
デル:HMM
適応音響モデル・
データ:HMM

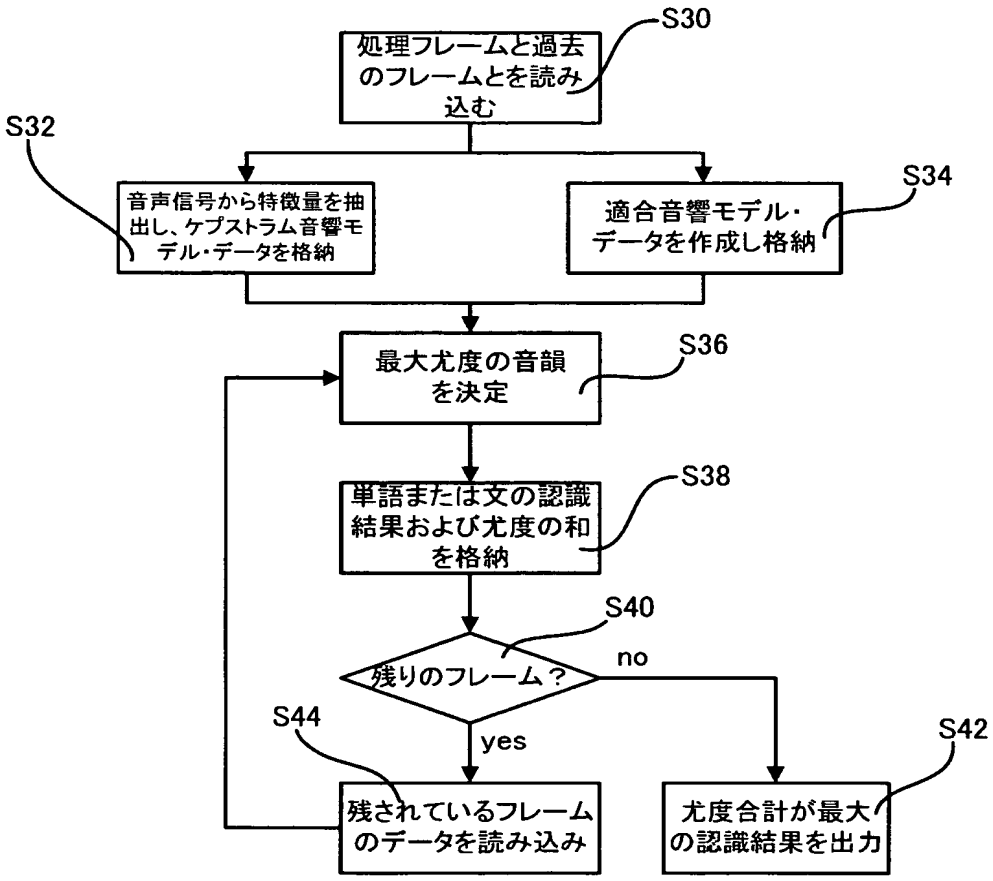
【図５】


26
28
24
言語モデル・データ格納部
適合音響モデル・データ格納部
音響モデル・データ格納部
出力結果
認識処理部
16
適合音響モデル・データ生成部
18
記憶領域
20
モデル・データ領域変換部
22
音声信号取得部
特徴量抽出部
14
12
10
音声信号

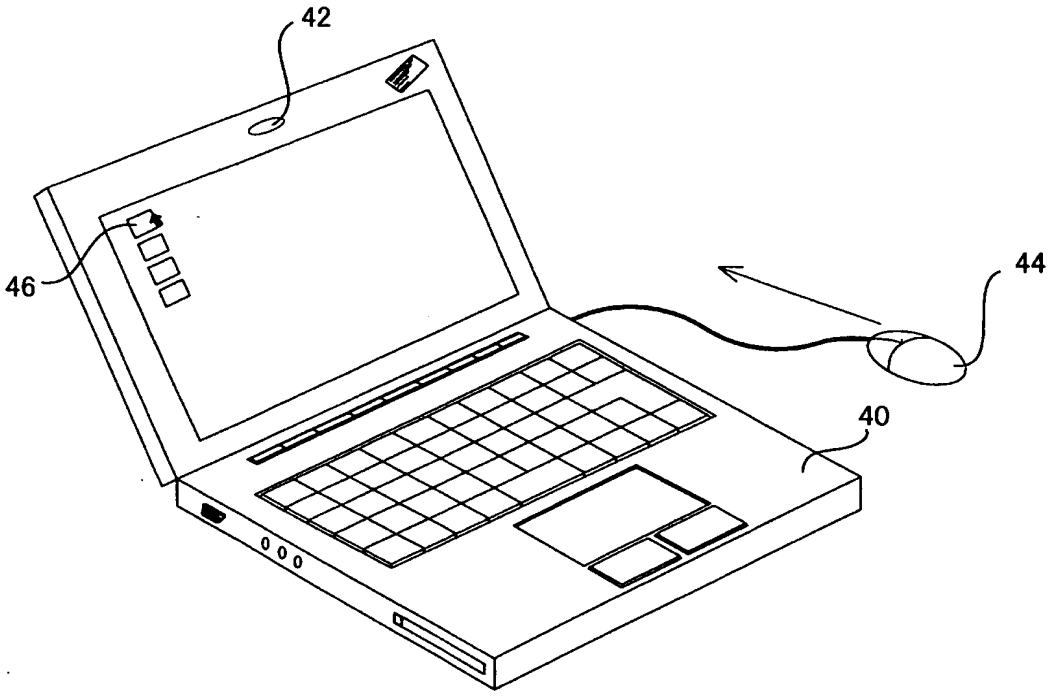
【図６】


18
過去の観測信号
30
バッファメモリ
36a
加算部
32b
モデル・データ領域変換部
28
適応音響モデル・データ格納部
34
残響予測係数算出部（EMアルゴリズム）
38
生成部
32a
モデル・データ領域変換部
36b
加算部
α×O（ω;tp）（過去観測信号）
24
音響モデル・データ格納部

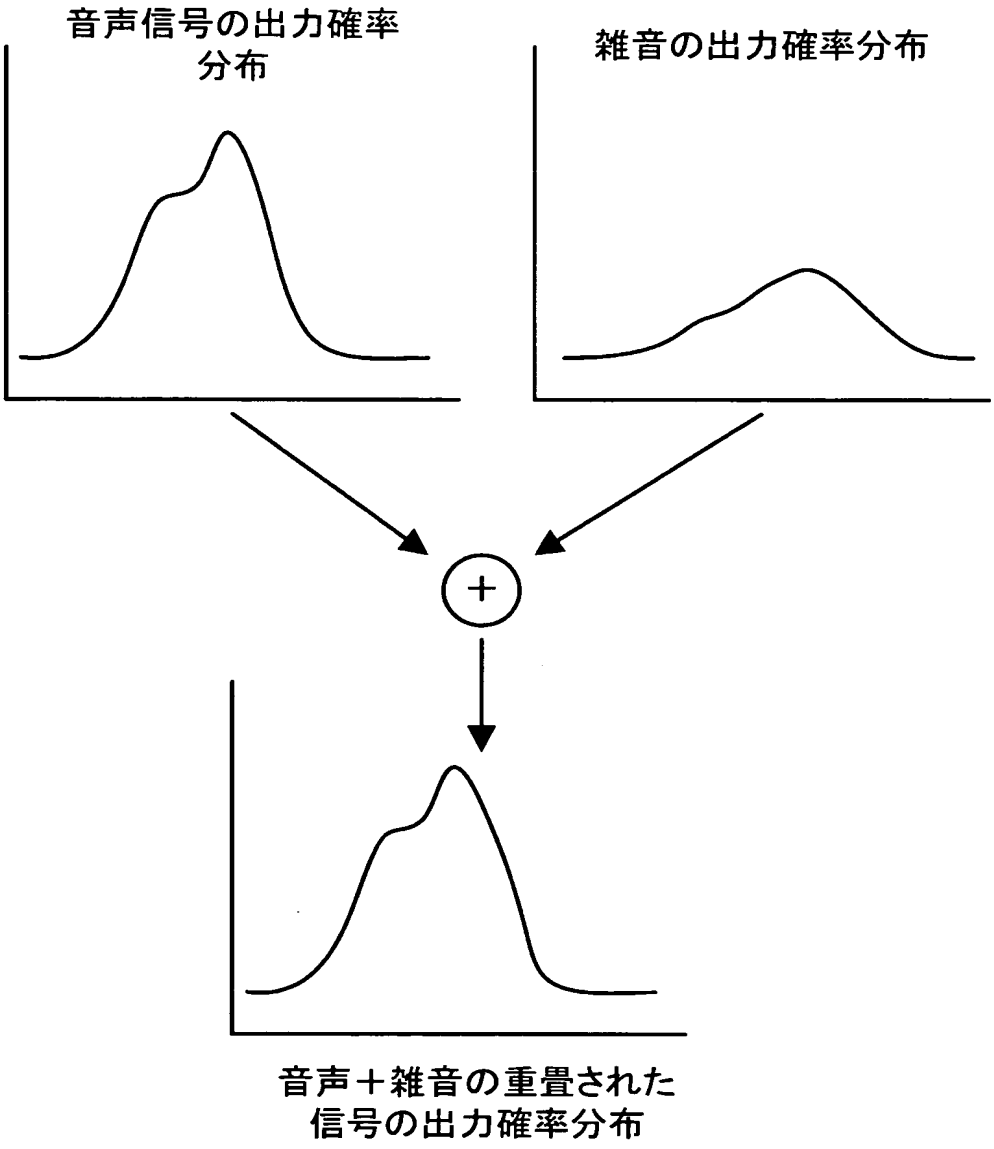
【図７】


処理フレームと過去のフレームとを読み込む
S30
S32
音声信号から特徴量を抽出し、ケプストラム音響モデル・データを格納
適合音響モデル・データを作成し格納
S34
最大尤度の音韻を決定
S36
単語または文の認識結果および尤度の和を格納
S38
S40
残りのフレーム？
no
yes
S44
S42
残されているフレームのデータを読み込み
尤度合計が最大の認識結果を出力

【図８】



【図９】



【書類名】　要約書

【要約】

【課題】　周囲環境からの残響がオリジナルの音声に重畳される場合であっても充分に、オリジナル音声を認識するための音声認識装置、音声認識方法、該制御方法をコンピュータに対して実行させるためのコンピュータ実行可能なプログラムおよび記憶媒体を提供する。

【解決手段】　コンピュータを含んで構成され音声を認識するための音声認識装置であって、該音声認識装置は、音声信号から得られる特徴量をフレームごとに格納する手段２０と、音響モデル・データおよび言語モデル・データを格納するための手段２４、２６と、その時点で処理するべき音声信号よりも前に取得された音声信号から残響音声モデル・データを生成し、残響音声モデル・データを使用して適合音響モデル・データを生成する手段１８と、特徴量と適合音響モデル・データと言語モデル・データとを参照して音声信号の音声認識結果を与える手段１６とを含む。

【選択図】　図５

認定・付加情報

特許出願の番号	特願２００３－１４３２２４
受付番号	５０３００８４２７０９
書類名	特許願
担当官	金井　邦仁　　３０７２
作成日	平成１５年　６月３０日

＜認定情報・付加情報＞

【特許出願人】
- 【識別番号】 390009531
- 【住所又は居所】 アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
- 【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】
- 【識別番号】 100086243
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内
- 【氏名又は名称】 坂口　博

【代理人】
- 【識別番号】 100091568
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内
- 【氏名又は名称】 市位　嘉宏

【代理人】
- 【識別番号】 100108501
- 【住所又は居所】 神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権
- 【氏名又は名称】 上野　剛史

【復代理人】 申請人
- 【識別番号】 100110607
- 【住所又は居所】 神奈川県大和市中央林間３丁目４番４号　サクライビル４階　間山国際特許事務所
- 【氏名又は名称】 間山　進也

出　願　人　履　歴　情　報

識別番号　　［３９０００９５３１］

１．変更年月日　　２０００年　５月１６日
［変更理由］　　名称変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション

２．変更年月日　　２００２年　６月　３日
［変更理由］　　住所変更
住　所　　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
氏　名　　インターナショナル・ビジネス・マシーンズ・コーポレーション